CATCH/IT: A Data Warehouse to Support Comprehensive Assessment for Tracking Community Health

Donald J. Berndt and Alan R. Hevner
College of Business Administration
University of South Florida

James Studnicki
College of Public Health
University of South Florida

ABSTRACT

A systematic methodology, Comprehensive Assessment for Tracking Community Health (CATCH), for analyzing the health status of communities has been under development at the University of South Florida since the early 1990s. CATCH draws 226 health status indicators from multiple data sources and uses an innovative comparative framework and weighted evaluation criteria to produce a rank-ordered list of community health problems. CATCH has been applied successfully in many Florida counties; focusing attention on high priority health issues and measuring the impact of health expenditures on community health status outcomes. Previously performed manually, we are using information technology (IT) to automate the CATCH methodology with a full-scale data warehouse, user-friendly forms and reports, and extended analysis and data mining capabilities. The automated system, CATCH/IT, will reduce the time to prepare community health status reports from months to days. In this paper, we present the current status of the project, along with the principal research and development issues and future directions of the project.

COMMUNITY HEALTH ASSESSMENTS

The United States now spends nearly a trillion dollars annually on health expenditures. Both as a percentage of national productivity and per capita, this amount is larger than any other nation in the world. However, this tremendous expenditure has not secured the U.S. a rank among the 'healthiest' nations. In fact, for many health indicators, such as infant mortality and measles immunizations, the U.S. ranks below some countries characterized as underdeveloped [1, 2]. It is noteworthy that the debate on national health care reform has dealt mostly with insurance coverage and medical care financing, while avoiding any serious discussion concerning the true health of the nation.

A major barrier to any discussion of the nation's health is that there is no uniform, accepted method for determining a community's, a state's, or a nation's health status. Although there are many sources of health data, there are no standard data definitions, formats, or reports across the health care industry. Thus, health care data is widely used (and misused) in an ad-hoc manner to justify managerial objectives of health institutions and agencies, a maze of mandated categorical funding, and a variety of political agendas.

The Institute of Medicine of the National Academy of Sciences, in its influential 1988 report on the Future of Public Health, emphasizes that health care assessment is one of the core functions of public health and recommends that there should be a regular and systematic collection, assemblage, and analysis of information on the health status and needs of communities [3]. A *community health profile* is made up of socio-demographic characteristics, health status and quality of life indicators, health risk factors, health resource indicators, and other measures which can be used to support priority setting, resource allocation decisions, and the evaluation of the impact of health programs. The intent of such a comprehensive profile is to help the community establish and maintain a broad strategic view of its health status and the various factors which influence it.

Since 1991, the College of Public Health at the University of South Florida has developed and enhanced a methodology for performing community health assessments; the Comprehensive Assessment for Tracking Community Health (CATCH). CATCH has been applied to 12 counties throughout Florida

1091—8280/98/$5.00

with great success. For each of the 226 health care indicators, the selected county is compared to the State average and an average based on three peer counties in the State. All indicators that are worse than both the State average and the peer counties' average are subjected to further analyses and prioritization. The final report provides a rank-ordered list of health indicators which represent a profile of the communities' most serious and challenging health problems. A more complete presentation of the CATCH methodology is found in [4, 5].

While the value of CATCH is incontrovertible, the ultimate deployment of CATCH throughout Florida and the nation is constrained by several serious limitations:

- The current process is labor-intensive and slow. Hundreds of individual sources of data must be identified and contacted. Data are often provided in hard copy formats and must be manually checked, validated, and entered into spreadsheets. With current methods, it takes 3 to 4 months to complete a CATCH report for one county.
- Longitudinal trend analyses over many years are cost prohibitive for most communities. Since each application is expensive and time-consuming, the capability to fund and produce annual assessments in a single community is limited.
- Most public health funding comes from state and federal governments. A state-wide CATCH assessment would help to prioritize funding and serve to enable effective program evaluation based on quantifiable outcomes assessment. Since nearly all data elements available in Florida are available in most other states, there is reason to be confident that CATCH might be expanded nationally and even internationally.
- With the massive amount of health care data involved, many interesting relationships and correlations of health status indicators can be found and investigated. Currently, in the manual system such discovery is not being done.

To overcome these limitations, we are building a data warehouse for the CATCH health care data. This data warehouse will dramatically reduce the time to produce a CATCH report for community assessments. The data warehouse will also provide a flexible infrastructure for ad-hoc exploration, data mining, and the customization of community reports. In this paper, we present the initial design for the data warehouse and the current progress on developing the user interfaces and data mining applications. Finally, we discuss the challenging problems and opportunities we face as the data warehouse continues to evolve.

CATCH DATA WAREHOUSE DESIGN

Important missions of a data warehouse include the support of decision-making activities and the creation of an infrastructure for ad-hoc exploration of very large collections of data. Decision makers should be able to pursue many of their investigations using browsing tools, without relying on database programmers to construct queries. The emphasis on empowering users places a premium on an understandable database schema that provides a natural basis for navigating through the data. The *star schema* or *dimensional model* has been recognized as a good structure for organizing many data warehouse components [6]. Figure 1 contains a fragment of the CATCH data warehouse schema, with simplified dimensions. The DEMOGRAPHIC table is one of the center fact tables (there is a family of such tables) and is the repository for items such as population levels. This fact table is surrounded by a collection of dimension tables: AGE, COUNTY, ETHNICITY, GENDER, and YEAR. Population levels are stored for each combination of age group, county, ethnic group, gender, and year. Several of these dimensions are part of larger hierarchies, such as COUNTY, STATE_REGION, STATE, NATL_REGION. This type of structure easily supports a wide range of queries. For instance, the population levels for a particular county can be calculated based on an arbitrary set of ethnic groups and age ranges.

An effective data warehouse structure must handle temporal information (i.e., historical data). This implies that the data warehouse will continue to grow in size as more and more data are collected. More importantly, the data warehouse must support temporal queries and the investigations of trends. This activity is crucial with regard to the CATCH methodology and health care data in general. The YEAR dimension table provides one mechanism for storing such historical data.

Data aggregation, especially with regard to dimensions such as geography and time, is a major issue in the data warehouse design. The CATCH methodology relies on 226 health status indicators drawn from an assembly of unrelated organizations with their own notions of the proper levels of data aggregation. In order to accommodate this varied data in the short-term, we define auxiliary fact tables which store data at coarse, pre-defined levels. For

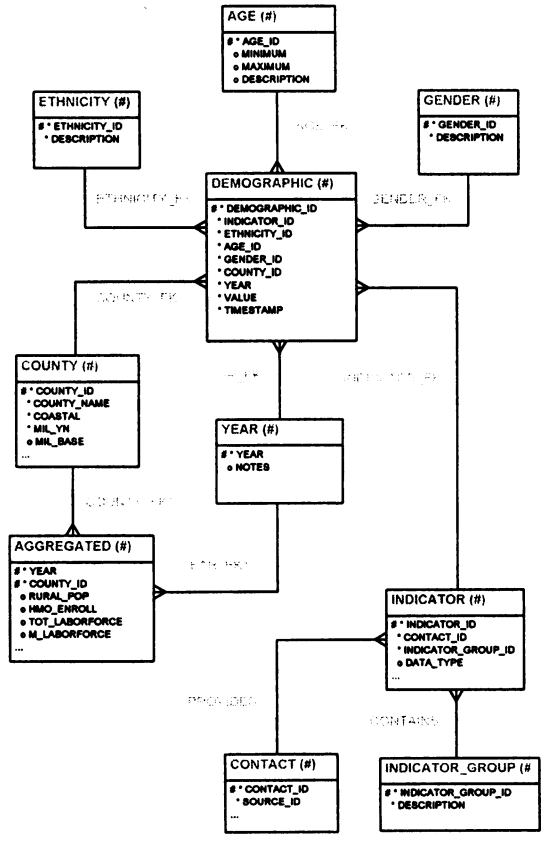

Figure 1

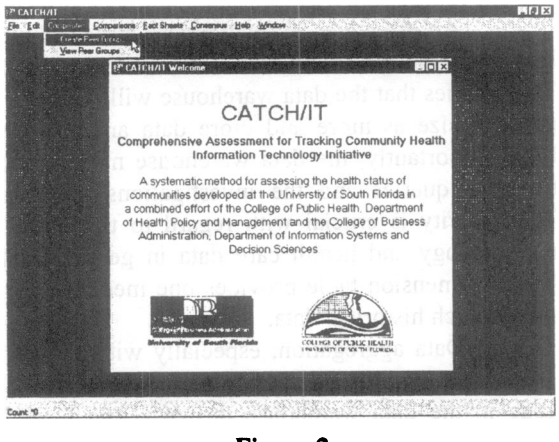

Figure 2

instance, the AGGREGATED table provides a repository for annual data collected by county. While this design provides the necessary information for generating a yearly county-level report, it lacks the flexibility to support interesting ad-hoc queries and data mining applications.

In addition to the fact and dimension tables, we also incorporate metadata within the data warehouse framework. As noted above, the CATCH methodology uses 226 health care indicators organized into 10 major indicator groups. The INDICATOR, INDICATOR_GROUP, and CONTACT tables are a subset of the metadata entities which are associated with the fact tables. This provides a useful mechanism for organizing the many indicators and tracking formats, reporting frequency, data sources, and descriptions. From the data warehouse administration perspective, the metadata support the large data collection effort, the generation of reports, and the ability to forecast data warehouse growth rates.

CATCH USER INTERFACE

In this section, several screens from the prototype CATCH data warehouse are presented with two purposes in mind. First, the screens provide a glimpse of the types of activities supported. Second, the screens provide a framework for describing some details of the CATCH methodology. The overall goal of the CATCH methodology is to provide a comparative report for a target county with respect to the state and a set of peer counties for the health care indicator groups. The main screen (Figure 2) provides access to target county selection, peer county selection, and reporting functions. Comparative reports are based on the 10 indicator groups, which include Demographic Characteristics, Socioeconomic Characteristics, Infectious Disease, Maternal and Child Health, Social and Mental Health, Physical Environmental Health, Sentinel Events, Health Status Indicators, Health Resource Availability, and Behavioral Risk Factors. The following steps outline the approach:

1. A target county is selected for investigation (Figure 3).
2. A set of peer counties is selected based on the similarity of population levels, income, ethnic composition, and other demographic indicators as depicted in Figure 3. One of the most important benefits of this data warehouse is that it provides a flexible infrastructure for modifying and refining peer selections in support of ad-hoc investigations. For example, custom reports that use alternate peer selections based on specific ethnic or age groups can be used to further explore particular health care issues.

3. A comparative screen for each major indicator group shows favorable/unfavorable ratings for individual indicators in a two-by-two matrix with respect to both the peer counties and the state. The comparative framework used for each indicator group is illustrated in Figure 4.

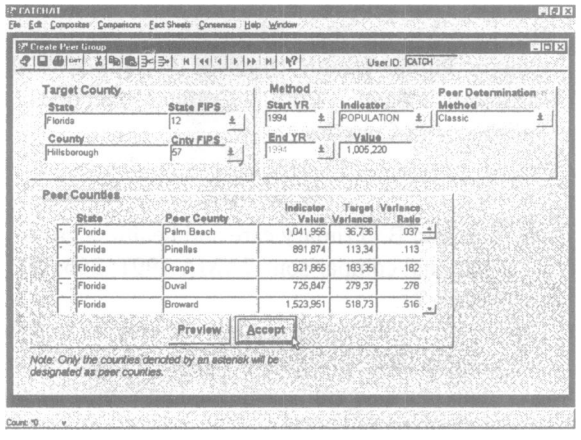

Figure 3

4. For selected indicators, more detailed analyses are performed. Factors, such as the number of people affected, the availability of efficacious treatment, economic impact, and trend analysis, are used to rank the indicators in order of priority. This rank order of indicators supports policy formulation and resource allocation decisions.

CATEGORY	INDICATOR	CO.	PEER	ST.
Socioeconomic	% Labor Force Unemployed	5.2	5.8	6.6
Maternal & Child Care	Infant Mortality	14.6	21.4	12.9
Infectious Disease	Tuberculosis Cases	0.31	0.57	0.41
Health Status	Colorectal Cancer	11.3	12.3	10.8
Resource Availability	Licensed Hospital Beds	5.7	4.5	4.2
Sentinel Events	Cervical Cancer, Late Stage	51.3	41.7	45.6
Physical/ Environmental	Drowning Fatalities	2.7	2.0	2.4
Social/Mental	Domestic Violence Cases	1027	864	1103
Behavioral Risk	Pap Smear	69.9	68.9	65.9

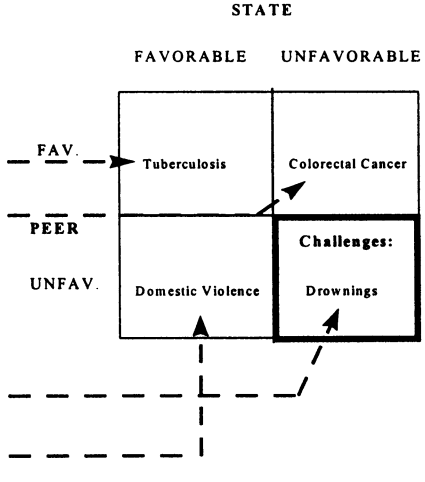

Figure 4

DATA WAREHOUSE DEVELOPMENT ISSUES

The development of a data warehouse is a complex endeavor and there are many issues that warrant further research and development. However, three issues stand out in the CATCH data warehouse project.

1. We are using an *evolutionary, incremental development approach* to data warehouse design and construction. This approach helps to reduce the risks in such a complex project, especially when operating with limited resources [7]. Our philosophy in the design of the CATCH data warehouse is to identify subsystems that can support useful tasks, foster familiarity with the system, and provide early evidence of success. This has led to an incremental development effort in which components have been defined and built with the goal of providing utility while evolutionary design and implementation initiatives proceed.
2. *Data availability* issues are of paramount concern. The CATCH methodology relies on a large set of health status indicators gathered from a diverse group of primary sources. One of the most difficult tasks is identifying a natural level of aggregation for the data (i.e., the temporal and geographic dimensions to be used) and pursuing the 'up-stream' negotiations in order to ensure that the data can be made available in the desired form.
3. Achieving high levels of *data quality* is essential. The CATCH data is intended to support health care policy formulation. Good, defensible policies require high quality data. We are investigating a number of approaches for certifying data quality. For example, the issue of quality is part of the 'up-stream' negotiations with primary data sources, data cleansing operations are integrated with load operations, automated data checks are incorporated within the warehouse, and redundant data provide a means for verification and recovery.

Currently, the prototype CATCH data warehouse is being constructed and initial data collection efforts are under way. Many of the health status indicators are available in aggregated form (i.e., by year and county). One of the important data collection tasks is to acquire critical data in a more fine-grained form using additional dimensions such as ETHNICITY and AGE. With respect to the data warehouse design, this effort involves migrating data from the AGGREGATED table to the family of dimensional fact tables. The data warehouse is being developed on both Unix and Windows NT platforms using the Oracle Enterprise Server. In addition, Oracle tools are being used to support the design and client-side development of forms and reports. The prototype CATCH/IT system will continue to evolve as it is used to support several planned county-based CATCH projects.

FUTURE DIRECTIONS

The future directions for the CATCH data warehouse cover both technical and health care related objectives. Incremental development of the warehouse will continue as all data indicators are integrated into the database and applications. In order to evaluate the CATCH/IT system, we will perform a pilot study for all Florida counties. This will lead the way for a comprehensive state-wide health care assessment. We are planning to develop Internet interfaces to provide widespread access to the data warehouse for research and policy formulation within external organizations. Lastly, the data warehouse provides an experimental infrastructure for innovative data mining techniques [8, 9]. The rich set of data, including demographic information, makes the data warehouse an excellent arena in which to apply data mining software with the potential of discovering important patterns in community health [10].

ACKNOWLEDGEMENTS

The prototype CATCH data warehouse is being constructed with the help of several College of Business Administration graduate students, including James Creed, Colleen Endres, Michael Morley, and James Nohelty. The College of Public Health research group directed by Dr. James Studnicki has been developing the CATCH methodology for several years. The members of that team include Barbara Myers and Barbara Steverson. The CATCH/IT initiative is truly an interdisciplinary effort and reflects the commitment of team members from both the College of Business Administration and the College of Public Health at the University of South Florida. This research is partially funded by the following organizations: Centers for Disease Control and Prevention; the Bert Fish Foundation, Inc.; Halifax-Fish Community Health; Memorial Health Systems, Inc.; Lee Memorial Health System; Florida Department of Health and Rehabilitative Services; Morton Plant Mease Health Care; and the County Public Health Units of Escambia, Walton, Okaloosa, Santa Rosa, Collier, Hendry, and Hillsborough Counties.

References

[1] Starfield B. Primary care and health: a cross-national comparison. JAMA 1991; 266(16): 2268-71.

[2] World Health Organization. The world health report 1995: bridging the gaps. Report of the Director-General. Geneva: The World Health Organization; 1995.

[3] Institute of Medicine. Summary of recommendations. In: Waterfall W, editor. The future of public health. Washington (DC): National Academy Press; 1988.

[4] Studnicki J. Evaluating the performance of public health agencies: information needs. American Journal of Preventive Medicine, Research, and Measurement in Public Health Practice 1995; 11(6): 74-80.

[5] Studnicki J, Steverson B, Myers B, Hevner A, Berndt D. Comprehensive assessment for tracking community health (CATCH). Best Practices and Benchmarking in Healthcare 1997; 2(5): 196-207.

[6] Kimball R. The data warehouse toolkit. New York: John Wiley & Sons; 1996.

[7] Corey M, Abbey M. Oracle data warehousing. New York: Oracle Press and Osborne McGraw-Hill; 1997.

[8] Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining. Menlo Park (CA): The AAAI Press; 1996.

[9] Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 1996; 39(11): 65-8.

[10] Berndt D. AX: searching for database regularities using concept networks. Proceedings of the Workshop on Information Technologies and Systems 1995; December: 142-51.